Computer-aided Document Indexing System

Mladen Kolar, Igor Vukmirović, Bojana Dalbelo Bašić and Jan Šnajder

Faculty of Electrical Engineering and Computing, University of Zagreb, Croatia

An enormous number of documents is being produced that have to be stored, searched and accessed. Document indexing represents an efficient way to tackle this problem. Contributing to the document indexing process, we developed the Computer-Aided Document Indexing System (CADIS) that applies controlled vocabulary keywords from the EUROVOC thesaurus. The main contribution of this paper is the introduction of the special CADIS internal data structure that copes with the morphological complexity of the Croatian language. CADIS internal data structure ensures efficient statistical analysis of input documents and quick visual feedback generation that helps indexing documents more quickly, accurately and uniformly than by manual indexing.

Keywords: information retrieval, document indexing, text statistics, EUROVOC.

1. Introduction

In the modern world overwhelmed by computers, huge storage spaces and widely accessible Internet, information is no longer a desired wealth. It is a cumbersome requirement that has to be dealt with efficiently.

Official documents have to be collected and neatly stored, at the same time providing quick access and search by means of indexing, a process of assigning keywords to documents. A kind of indexing is keyword assignment, identification of appropriate keywords from the controlled vocabulary of a reference list (a thesaurus) [10]. Lacking good automated systems, one had to go through every single legal document and assign the keywords manually. Obviously, the need existed to relieve human indexers of this long, tedious and, above all, expensive work.

Challenged by this, we started designing and implementing a Computer-Aided Document Indexing System (CADIS), presented here.

The main motive to develop the system was the fact that official documents have to be indexed according to multilingual EUROVOC thesaurus. It covers the fields in which the European Communities are active. Its form of a controlled and structured list provides a means of indexing the documents in the documentation systems of the European institutions and of their users [2]. It organizes over 6,000 descriptors (classes) from 21 different fields (e.g. politics, finance, science, social questions, organizations, foodstuff, etc.) hierarchically into a maximum of 8 levels.

Due to the multilingual nature of EUROVOC, it is very suitable for applications such as cross-lingual document similarity calculation, multilingual clustering and categorization, and cross-lingual document retrieval and information access [10].

From the beginning, CADIS was designed with EUROVOC in mind, but not limited to it, making it possible to use the statistical output of the system for other purposes too.

CADIS does not perform automatic document indexing, it helps humans to make the process of standard intellectual indexing easier, by providing results of built-in statistical and natural language processing techniques. CADIS speeds up indexing and ensures building a set of uniformly indexed documents.

The paper is organized as follows. Section 2 gives an insight into related work. The functionality of CADIS and some examples of use are given in Section 3. Section 4 describes

abstraction of the object model and problems tackled during implementation, while Section 5 describes multilingual aspects of the system. In Section 6 conclusion and future work are given.

2. Related Work

Automatic document indexing (or classification) is an old problem, already described in 1961. Maron examined a technique for classifying documents automatically, according to their subject contents [7].

Automated document indexing using EUROVOC thesaurus is a problem that occupies many scientists. In 1997, Ferber [4] built an application which used a multilingual thesaurus for the retrieval of English documents using search terms in other non-English languages. A more recent example of such application was built by Steinberger and Pouliquen who introduced statistical methods for cross-lingual indexing [10, 13, 14]. An automatic indexing system that could work in a language- independent environment was developed. JRC workshop on EUROVOC [5] brought some ideas and progress on automated indexing.

Montejo-Raez in his papers [8] describes the use of information retrieval for indexing techniques. Our own approach to the problem is multi-monolingual. The CADIS system can deal with several languages, but the results will always be displayed in the same language as the text. It does not support automated indexing, due to the lack of a substantial learning set of pre-indexed documents in Croatian. Its purpose is to help human indexers, using statistical processing on text, to index documents that will form a learning set for building an automated document indexer.

3. Functionality

3.1. System Input

XML [3], the selected format for information interchange has many favorable characteristics for this system. As it is an industry standard, there are good reasons to believe that some of the input will already comply with this format. On the other hand, existing documents in other formats (HTML, PDF etc.) can be converted to XML in a more or less simple procedure. Furthermore, XML enables the coding of additional information to the text itself, such as title, source, paragraphs, formatting etc. needed for visual reconstruction of documents within the CADIS.

From the document received as the input, an internal data structure, representing the XML document, is built and a sequence of statistical and lexical analyses is carried out.

3.2. Statistics and Visual Feedback

An example of the CADIS user interface is shown in Fig. 1. After the initial processing, a visual representation of the XML document currently being indexed is shown on the left hand side of the user interface. The visual representation is generated by the PEI program in RTF format [11] using formatting and paragraph information provided in the XML document.

Primarily intended for documents in Croatian, during the design of CADIS special attention had to be paid to the morphological complexity of the language. Certain Croatian words, such as nouns and adjectives, can have a number of different morphological forms, depending on the number, gender, degree etc. However, statistically, each of these forms is treated as one occurrence of the word in its basic form, also called the lemma [16].

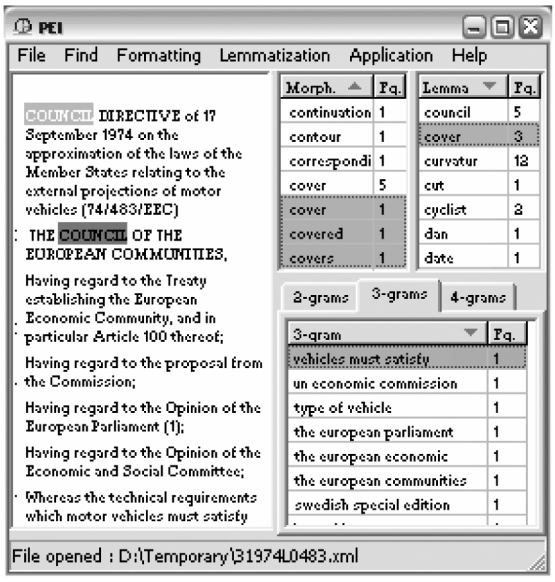

Fig. 1. Example of the CADIS user interface.

Lemmatization process was implemented for Croatian and English. Automatically acquired morphological lexica was used for Croatian [15], for lemmatization of English Porter's stemming algorithm was used [9].

The output of the statistical analysis of the text consists of two lists: the list with all morphological forms found in the document with their frequencies and the list that contains only lemmas of these words and their frequencies. Both of these lists enable sorting according to words or frequencies and locating the occurrences of any word or lemma within the document by a double-click. The same functionality is available throughout the menu.

The statistical analysis of the text provides a solid ground for the visual feedback mentioned above. With each search a red background coloring visually emphasizes all the occurrences of the sought word or lemma, while the selection jumps from one occurrence to the other. The system also enables visual emphasizing of the most frequent lemmas in the document (an example shown in Fig. 2.), as well as hiding irrelevant information from the text, such as stop-words, words with minimal semantic meaning (e.g. conjunctions etc.) [12]. This makes it easier for the indexer to browse through the document, paying attention only to statistically relevant parts of the document.

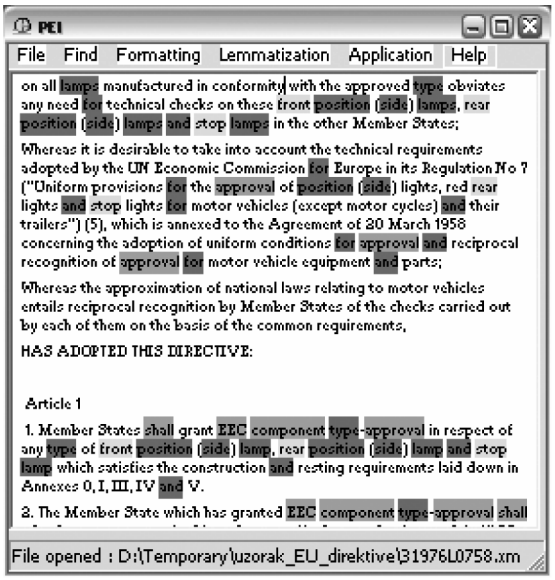

Fig. 2. Visual feedback example.

3.3. N-grams

Another relevant information for document indexation is the occurrences of N-grams. An N-gram is a sequence of N tokens. We define token as a word delimited by space or a single punctuation mark. Especially interesting N-grams are those sequences of words that tend to appear next to each other in a text more often than what we would consider random. Such occurrences of two or more words that we use as a conventional way to say things are called *collocations* [1]. Collocations carry much information that is usually different from information we would gain when looking at the words separately.

We could say that collocations have added meaning to the combination of words. Idioms and phrases are the most extreme examples of such added meaning. For example, *machine* and *learning* have different meanings when looked separately from the collocation *machine learning*.

CADIS provides statistical information on collocation. Lists with 2-grams, 3-grams or 4-grams and their frequencies are presented to the user. All of these lists can be sorted according to collocations or their frequencies. Only N-grams appearing in the corpus and identified as collocations are shown. The input to our system consists only of collocations that were found using statistical methods over the corpus.

Collocations can appear in different morphological forms, but all of them are mapped to one in the list. The most frequent morphological form of a collocation represents a group of collocations in the list with their frequencies summed up.

4. Implementation

4.1. CADIS Internal Data Structure

Displaying documents in CADIS requires XML document format. Forcing the user to convert different document formats to XML might seem an extra step in displaying the document. For example, if the original document was in HTML format, we could convert the document internally in the RTF format and display it in the original, HTML format.

This extra effort is needed to keep the visual feedback under control as much as possible. Some implemented functions would require modifying certain text attributes of the original document, so generating the entire RTF inside the system is faster and more accurate.

Internal representation of the input document is very important for the efficiency of the system. Setting the speed of traversal and low memory consumptions as priorities, we designed the CADIS internal data structure of the document that serves as the base of the whole system. A simple XML document is shown in Fig. 3 and graphic abstraction of the internal data structure representing the document is shown in Fig. 4.

The XML format itself is represented as a standard first child — next sibling tree, the root node representing the root tag of the document. The leaf nodes on the lowest level, representing the very text of the document, are of great interest. Instead of keeping the words, punctuation etc. in the nodes of the tree, a table is created of all the morphological forms found in the document. The leaf nodes point to records in this table, saving space due to frequently appearing words. During the creation of the tree, the number of occurrences of each form is being counted.

```
<head>
  <title>To love!</title>
  <body>To be loved!</body>
</head>
```

Fig. 3. An XML input example.

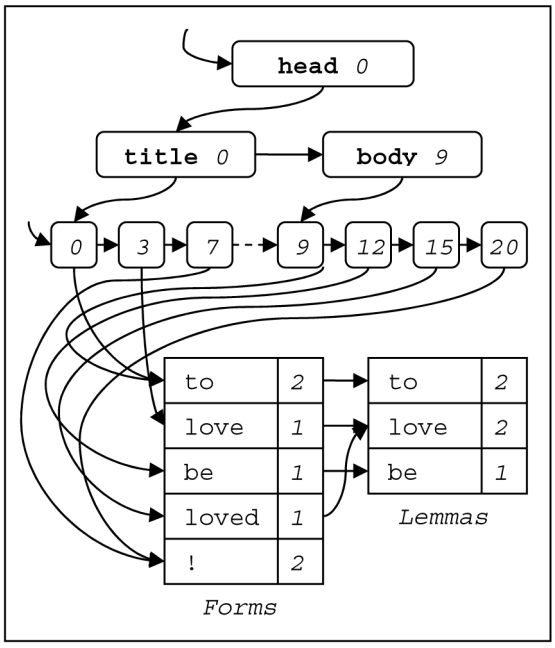

Fig. 4. Abstraction of CADIS internal data structure built from the XML example in Fig. 3.

4.2. Benefits of the CADIS Data Structure

Due to morphological complexity of the language, a quick, but efficient way of retrieving lemmas of every word encountered in the document had to be designed. This information is required for the basic statistical function — the lemma count. The straightforward way was to build a database of a substantial number of lemmas of Croatian words together with all the morphological forms. The goal was accomplished by executing a simple string search through this database.

Lemmas are kept in a similar table. Each word found in the document points to its lemma in this table. As shown in the example, different morphological forms of the same word point to the same lemma. Thus, we know that the lemma “love” occurred twice in the document, although only once actually in the form “love”.

The remaining features to be explained are crucial for the speed of generating the visual feedback in the form of RTF format. Firstly, all leaves in the tree are connected in one linked list. This enables an easy traversal search of the nodes representing the very text of the document, ignoring for a moment the XML structure. Furthermore, for every element of the structure (tag, word, punctuation etc.) the information of its distance from the beginning of the document is kept. Due to specific characteristics of the programming environment, the word at a certain distance from the beginning of the document has to be found quickly. This redundant information in the structure makes time complexity of this search $O(d^*n)$, where d is an average depth of XML tags, and n is an average number of tokens in one XML tag. These features are also depicted in Fig. 2.

4.3. Capturing N-grams

N-grams do not necessarily consist of continuous tokens. N-grams at a distance are captured by the collocation window. Suppose that we have a collocation window of size M and we are capturing N-grams of size N. We combine every N tokens from this window to create N-grams. Because of the morphological complexity of the language, one collocation can be represented by two or more different N-grams. When captured, each token in the N-gram is replaced by a set of its lemmatized forms.

Due to outer homography (the case when two different words have the same word-form) each word can be lemmatized to several lemmas. If each set of a lemmatized N-gram contains a word from the collocation retrieved from the corpus, it is considered as an occurrence of the collocation. Collocations found are stored in a container, and with each repetition, its frequency is increased.

For capturing N-grams, modification of algorithm given by Mladenic et al. [6] was implemented. Modifications were necessary due to the homography described above. N-grams are captured by traversing nodes in a linked list. Time complexity of search for all collocations is $O(l^*m)$, where l represents the length of text, and m represents the number of collocations in the corpus, assuming that two strings can be compared in constant time.

5. Multilingual Aspects of CADIS

Document indexing using the EUROVOC thesaurus is a continuous task for all member countries of the European Union, as well as those aspiring for membership. Knowing that no similar system was designed previously and that each of these countries would benefit significantly from it, CADIS is designed to be fully functional multilingually, in three different aspects.

Firstly, the EUROVOC itself is a multilingual thesaurus, published currently in 16 official languages of the European Union and a number of other European languages.

Secondly, input documents can be written in a number of different languages. The key problem here lies in the fact that words in the documents have to be lemmatized in order for the system to function properly. The process of lemmatization is completely language-specific, but CADIS could be easily modified to accept inputs from any language. The lemmatization module keeps the vocabulary of the language, as well as the list of stop words in an extern file. This file can be generated for any language, thus making it possible for the system to work with

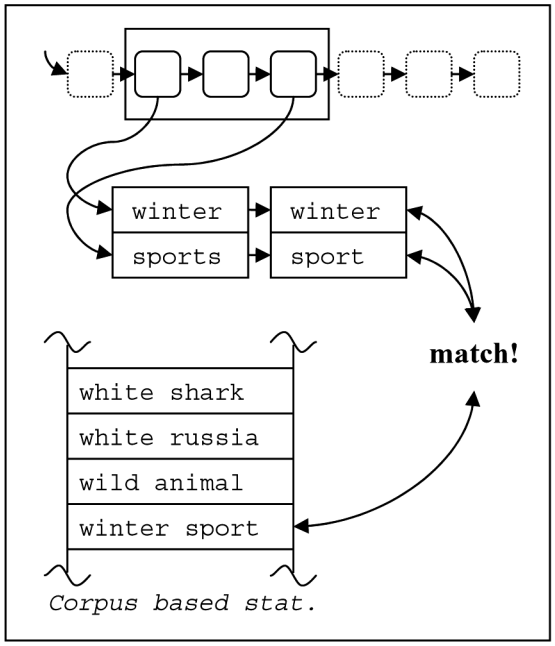

Fig. 5. Graphic abstraction of capturing an occurrence of collocation “winter sport”.

Fig. 6. Multilingual aspect of CADIS — Croatian user interface indexing a Croatian document.

documents in different languages without modifying the inner structure. To date, this concept has been implemented for Croatian and English [9, 15].

Finally, the third aspect of interest is the user interface. In order to make CADIS easily adaptable to different languages, specific elements of the user interface system, such as captions, labels, user messages etc., are kept in an extern file.

6. Conclusion and Future Work

Computer-Aided Document Indexing System (CADIS) that provides different statistical and visual information, supporting an efficient and uniform indexing process is described in the paper.

Some problems tackled during the design and implementation of the CADIS include the morphological complexity of Croatian language, strongly connected with implemented statistical functions of word and lemma count and collocation finding.

Space and speed efficiency demanded implementation of a proprietary data structure.

Multilingual aspects of the system include a multilingual thesaurus, a simple user interface translation and an independent lemmatization algorithm.

Several independent evaluations showed that the process of automatic descriptor assignment does not reach the quality of human indexation [13]. Even the existing automatic indexing systems [4, 10] serve mainly as a first step in the process of keyword assignment, offering descriptors' suggestions which will then be verified by a human indexer. Having this in mind, CADIS comes close to the functionality of an automatic system and is especially valuable in circumstances where no machine learning methods are applicable due to the lack of a learning set.

Ultimately, CADIS will help to generate a substantial number of documents indexed by EUROVOC descriptors in a more uniform way, at the same time saving human and financial resources. The obtained set of indexed documents will then be used to train automatic descriptor assignment systems using machine learning methods. Automatically found EUROVOC descriptors will be included as an extra option to CADIS.

7. Acknowledgments

We would like to thank the team from the Faculty of Electrical Engineering and Computing, Croatian Information Documentation Referral Agency, and the Department of Linguistics, Faculty of Philosophy for outstanding work on the project Text Mining System (2003–082) funded by the Ministry of Science, Education and Sports of the Republic of Croatia.

References

[1] BRIDGET T. MCINNES, Extending the Log Likelihood Measure to Improve Collocation Identification, Master thesis, University of Minnesota, December 2004.

[2] EUROVOC thesaurus. European Union publications office. `http://europa.eu.int/celex/eurovoc/` [02/12/2005].

[3] Extensible Markup Language (XML). Version 1.0 Specification. World Wide Web Consortium, 2004 `http://www.w3.org/TR/2004/REC-xml-20040204/` [02/13/2005].

[4] R. FERBER, Automated Indexing with Thesaurus Descriptors: A Co-occurrence based Approach to Multilingual Retrieval. In: Research and Advanced Technology for Digital Libraires. *Proceedings of European Conference of Digital Libraries* 1997.

[5] Joint Research Center, Workshop on EUROVOC - Addressing the Language Barrier Problem `http://www.jrc.cec.eu.int/langtech/Eurovoc/Eurovoc-Workshop_Sept2004.html` [02/22/2005].

[6] D. MLADENIC, M. GROBELNIK, (1998) Word sequences as features in text-learning. *Proceedings of the Seventh Electrotechnical and Computer Sc. Conference* ERK'98, Ljubljana, Slovenia: IEEE section, 1998, pp. 145–148.

[7] M. E. MARON, Automatic indexing: An Experimental inquiry. In: *Journal of the ACM*, Volume 8, 1961.

[8] A. MONTEJO-RAEZ, Toward conceptual indexing using automatic assignment of descriptors. In: Proceedings of the AM 2002 Workshop on Personalization Techniques in Electronig Publishing, Malaga, Spain, May 2002.

[9] M. F. PORTER, An algorithm for suffix stripping. In: Jones S, Willet K, Willet P. Readings in *Information Retrieval*, 1997.

[10] B. POULIQUEN, R. STEINBERGER, C. IGNAT, Automatic Annotation of Multilingual Text Collections with a Conceptual Thesaurus. In: *Proceedings of the Workshop Ontologies and Information Extraction Summer School The Semantic Web and Language Technology — Its Potential and Practicalities*. Bucharest, Romania, 28. July — 8. August 2003.

[11] Rich Text Format (RTF) Specification. Microsoft Corporation, 1999. `http://msdn.microsoft.com/library/default.asp?url=/library/en-us/dnrtfspec/html/rtfspec.asp`.

[12] T. SKOPAL, ACB Compression Method and Query Preprocessing in Text Retrieval Systems. In: *Proceedings of DATESO* 2002, Desna, Czech Republic.

[13] R. STEINBERGER, Cross-lingual Keyword Assignment. In: *Proceedings of the XVII Conference of the Spanish Society for Natural Language Processing*, Jaen, Spain, 12–14 September 2001.

[14] R. STEINBERGER, B. POULIQUEN, Cross-lingual Indexing. Final Report for the IPSC Exploratory Research Project. *JRC Internal Note*, October 2003.

[15] J. ŠNAJDER, Rule-based automatic acquisition of large-coverage morphological lexicons for information retrieval. Tech. Report, MZOŠ. 2003–082, ZEMRIS, FER, University of Zagreb; 2005.

[16] M. TADIĆ, B. BEKAVAC, Preparation of POS tagging of Croatian using ClaRK System. In: *Proceedings of RANLP2003 Conference*, Borovtes, Bulgaria, 2003.

Recived: June, 2005.
Accepted: October, 2005.

Contact address:
Mladen Kolar
Faculty of Electrical Engineering and Computing
University of Zagreb
Unska 3
10000 Zagreb
Croatia
`mladen.kolar@fer.hr`

Igor Vukomirović
Faculty of Electrical Engineering and Computing
University of Zagreb
Unska 3
10000 Zagreb
Croatia
`igor.vukomirovic@fer.hr`

Bojana Dalbelo Bašić
Faculty of Electrical Engineering and Computing
University of Zagreb
Unska 3
10000 Zagreb
Croatia
`bojana.dalbelo@fer.hr`

Jan Šnajder
Faculty of Electrical Engineering and Computing
University of Zagreb
Unska 3
10000 Zagreb
Croatia
`jan.snajder@fer.hr`

MLADEN KOLAR is a fifth year student of computing at the Faculty of Electrical Engineering and Computing, University of Zagreb. He is currently writing a diploma thesis at the Department of Electronics, Microelectronics, Computer and Intelligent Systems at the same faculty. His research interests lay in the field of machine learning and its application on text analysis.

IGOR VUKMIROVIĆ is a fifth year student of computing at the Faculty of Electrical Engineering and Computing, University of Zagreb. He is currently writing a diploma thesis at the Department of Electronics, Microelectronics, Computer and Intelligent Systems at the same faculty. His research interests lay in the field of machine learning, especially in Hidden Markov Models and its application in bioinformatics.

BOJANA DALBELO BAŠIĆ received her B.Sc. degree in mathematics from the Faculty of Sciences, University of Zagreb, in 1982 and the M.Sc. and PhD degrees in computer science from the Faculty of Electrical Engineering and Computing, University of Zagreb, in 1993 and 1997, respectively. She has published more than forty papers, mostly concerning applied multivariate statistics, fuzzy temporal reasoning and intelligent systems. Currently, she is associate professor at the Faculty of Electrical Engineering and Computing, University of Zagreb. Her currrent research interests include machine learning and applications to data and text mining.

JAN ŠNAJDER received his B.Sc. degree in computing from the Faculty of Electrical Engineering and Computing, University of Zagreb, in 2002. He is currently a PhD student at the Department of Electronics, Microelectronics, Computer and Intelligent Systems at the same faculty. His research interests lie in the field of aritificial intelligence, in particular multiagent systems, machine learning and natural language processing.